Causes

Rothman, K. American Journal of Epidemiology 104:6; 587-592

PM3003572244

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

AMERICAN
Journal of Epidemiology
Formerly AMERICAN JOURNAL OF HYGIENE
© 1976 by The Johns Hopkins University School of Hygiene and Public Health
VOL. 104 DECEMBER, 1976 NO. 6

Reviews and Commentary

CAUSES

KENNETH J. ROTHMAN

The conceptual framework for causes presented here is intended neither as a review nor an expansion of knowledge, but rather as a viewpoint which bridges the gap between metaphysical notions of cause and basic epidemiologic parameters. The focus, then, is neither metaphysics nor epidemiology, but the gulf between them. In the same spirit as recent discussion on these pages about definitions of basic epidemiologic terms such as *rate* (1), common agreement on the conceptual interrelationship of causes may facilitate communication about causes of illness.

A strong motivation for presenting this scheme is the often-heard confusion of two important but distinct epidemiologic issues: confounding and effect modification. These two properties of variables have different areas of relevance (2). The confounding property is not an intrinsic characteristic of any variable. Confounding, defined as distortion in an effect measure introduced by an extraneous variate, occurs only in the context of a particular study, and the same variable which confounds in one study may not confound the same association in another study setting. In fact, the principles of good study design may call for preventing a potentially confounding variable from being confounding —for example, by matching in the selection of subjects. On the other hand, effect modification, defined as differing values of the effect measure at different levels of another variate, is an inherent characteristic of the relationship between two causes of an illness (effect). This relationship is not governed by the particulars of any study; it is an unalterable fact of nature. To be sure, the magnitude of effect modification depends on the scale of measurement of the effect—for example, a risk ratio which is constant over age generally implies a risk difference which changes with age. But it has been proposed that the risk difference scale is the appropriate scale for assessing effect modification, other scales representing metameters of the risk difference which distort to varying degrees the assessment of effect modification (3). Effect modifiers may or may not be confounders in a given study, or they may confound in some studies but not in others. Conversely, confounders may or may not be effect modifiers. The following discussion presents a scheme for the interrelationship of causes which may provide a useful way for thinking about effect modification as a description of nature.

Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.

PM3003572245

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

Types of causes

A *cause* is an act or event or a state of nature which initiates or permits, alone or in conjunction with other causes, a sequence of events resulting in an *effect*. A cause which inevitably produces the effect is *sufficient*. The inevitability of disease after a sufficient cause calls for qualification: disease usually requires time to become manifest, and during this gestation, while disease may no longer be preventable, it might be fortuitously cured, or death might intervene.

Common usage makes no distinction between that constellation of phenomena which constitutes a sufficient cause and the components of the constellation which are likewise referred to as "causes". Another qualification for sufficient causes is restriction to the minimum number of required component causes; this implies that the lack of any component cause renders the remaining component causes insufficient. Thus, measles virus is referred to as the cause of measles, whereas a sufficient cause for contracting measles involves lack of immunity to measles virus and possibly other factors in addition to exposure to measles virus. The term *cause*, then, does not specify whether the reference is to a sufficient cause or to a component of a sufficient cause.

Most causes that are of interest in the health field are components of sufficient causes, but are not sufficient in themselves. Drinking contaminated water is not sufficient to produce cholera, and smoking is not sufficient to produce lung cancer, but both of these are components of sufficient causes. Identification of all the components of a given sufficient cause is unnecessary for prevention, in that blocking the causal role of but one component of a sufficient cause renders the joint action of the other components insufficient, and prevents the effect. Even without being able to identify the other components of the sufficient cause for lung cancer, of which smoking is one component, it is possible to prevent those cases of lung cancer which would result from that sufficient cause by removing smoking from the constellation of components.

A specific effect may result from a variety of different sufficient causes. The different constellations of component causes which produce the effect may or may not have common elements. If there exists a component cause which is a member of every sufficient cause, such a component is termed a *necessary* cause. Necessary causes are often identifiable as part of the definition of effect. For example, the possession of a vermiform appendix is necessary for appendicitis, and infection with the tubercle bacillus is a necessary cause for tuberculosis. Though sometimes devoid of useful significance, a necessary cause can be a useful component cause to identify. Whereas many different component causes have been identified for several types of cancer, the hope exists for identification of a final common pathway representing a necessary cause for cancer of all types.

Figure 1 is a schematic illustration of the causal components of a disease. The disease (effect) has three sufficient causal complexes, each having five component causes. In this scheme "A" is a necessary cause, since it appears as a member of each sufficient cause. On the other hand, the component causes "B", "C" and "F", which each appear in more than one sufficient cause, are not necessary causes, because they fail to appear in all three sufficient causes.

Etiologic fraction

Causal research of disease focuses on components of sufficient causes, whether necessary or not. The public health importance of a component cause of disease in a particular population is determined by the fraction of the disease (the effect) which results from the sufficient causes(s) to which the component cause belongs. The epidemiologic parameter *etiologic fraction*

PM3003572246

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

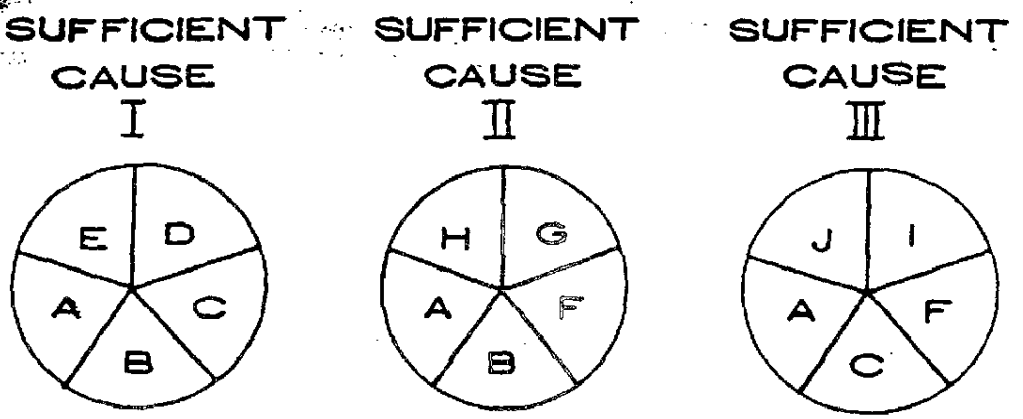

FIGURE 1. Conceptual scheme for the causes of a hypothetical disease.

(4) (population attributable risk) measures this dimension of a cause-effect relationship. Each component of a sufficient cause has, as its etiologic fraction, the fraction of disease attributable to that sufficient cause (plus the fraction attributable to any other sufficient causes which contain the same component). Consider, for example, the causes schematically represented in figure 1. If Sufficient Cause I accounts for 50 per cent of a disease, Sufficient Cause II 30 per cent, and Sufficient Cause III 20 per cent, the etiologic fractions for each of the component causes are: A, 100 per cent; B, 80 per cent; C, 70 per cent; D, 50 per cent; E, 50 per cent; F, 50 per cent; G, 30 per cent; H, 30 per cent; I, 20 per cent; and J, 20 per cent. (The example would have to be modified slightly if there were individuals with more than one sufficient cause because for these people blocking one sufficient cause would not prevent disease.) This example illustrates that the sum of etiologic fractions of a set of factors is not limited above by unity, as is observed with alcohol and tobacco as etiologic agents for mouth cancer, each having an etiologic fraction greater than 50 per cent.

RISK

The notion of a risk as a continuous measure obscures the concept of risk as applied to an individual. For an individual, risk for disease properly defined takes on only two values: zero and unity. The application of some intermediate value for risk to an individual is only a means of estimating the individual's risk by the mean risk of many other presumably similar individuals. The actual risk for an individual is a matter of whether or not a sufficient cause has been or will be formed, whereas the mean risk for a group indicates the proportion of individuals for whom sufficient causes are formed. An individual's risk can be viewed as a probability statement about the likelihood of a sufficient cause for disease existing within the appropriate time frame.

STRENGTH OF A CAUSAL RISK FACTOR

The model proposed also illustrates how characterization of risk factors as "strong" or "weak" has no universal basis. A component cause which requires, to complete the sufficient cause, other components with low prevalence is thereby a "weak" (component) cause. The presence of such a component cause modifies the probability of the outcome only slightly, from zero to an average value just slightly greater than zero, reflecting the rarity of the complementary component causes. On the other hand, a component cause which requires,

PM3003572247

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

to complete the sufficient cause, other components which are nearly ubiquitous is a "strong" (component) cause. In epidemiologic terms, a weak cause confers only a small increment in disease risk, whereas a strong cause will increase disease risk substantially.

Thus the strength of a causal risk factor depends on the prevalence of the complementary component causes in the same sufficient cause. But this prevalence is often a matter of custom, circumstance or chance, and is not a scientifically generalizable characteristic. Consider the following simplified example (5): in a society where most people eat high phenylalanine diets, inheritance of the (rare) gene for PKU would appear to be a "strong" risk factor for phenylketonuric mental retardation, and phenylalanine in the diet would appear to be a weak risk factor. In another society, however, in which the gene for PKU is very common and few people eat high phenylalanine diets, inheritance of the gene would be a weak risk factor and phenylalanine in the diet would be a strong risk factor. Thus, the strength of a causal risk factor, as it might be measured by the "risk ratio" (relative risk) parameter, is dependent on the distribution in the population of the other causal factors in the same sufficient cause. The term *strength* of a causal risk factor retains some meaning as a description of the public health importance of a factor. However, the common epidemiologic parlance about strength of causal risk factors is devoid of meaning in the biologic description of disease etiology.

Synergy

Synergy, also termed effect modification or positive interaction, may be defined as the relationship between factors which exhibit a joint effect that exceeds the sum of the separate effects, with the effects measured on an appropriate scale (3,6). Synergy implies that two component causes are members of the same sufficient cause. Neither of two such causal components of a sufficient cause can have any effect (as part of that sufficient cause) without the presence of the other causal component. Two such causes, and, indeed, all the components of a given sufficient cause, are mutually synergistic. Thus, inheritance of the PKU gene and phenylalanine in the diet are synergistic in producing phenylketonuric mental retardation. If two causes are components of different sufficient causes for the same effect, and are not mutual members of any other sufficient cause for that effect, they will exhibit no synergy and are thus considered *independent* in the biologic (not statistical) sense. If two components of a sufficient cause have no effect outside that sufficient cause, they will exhibit complete synergy, in the sense that no increase in risk can occur from either factor unless both factors are present. It is commonly observed that causes seem to have less than complete synergy with other causes, because some increase in the probability for the effect is observed even when the causes occur in the absence of complementary, synergistic causes. This pattern results from such causes being members of more than one sufficient cause. Synergy results from component causes being mutual members of a sufficient cause, but the component causes each may also be members of other sufficient causes with different complements, thereby also having independent effects and making the overall interrelationship of two causes one of incomplete synergy. Thus, two factors may be mutual members of a sufficient cause, and each may separately be a member of another sufficient cause. Each factor in the absence of the other has an effect through one sufficient cause, but together they have an effect through three sufficient causes, displaying incomplete synergy. Necessary causes are at least partially synergistic with all other causes of the same effect, inasmuch as a necessary cause is a member of every sufficient cause.

Figure 1 suggests many synergistic relationships. For example, "D" and "E" are

PM3003572248

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

completely synergistic with each other and each is partially synergistic with "A", "B" and "C". Partial synergy exists between "B" and "C"—their effect is dependent on their joint presence in one sufficient cause, but each also has independent effects in another sufficient cause. The extent to which two factors are synergistic depends, like the strength of a causal risk factor, on the distribution of other factors in a particular population. Consider synergy between factors "B" and "F" in figure 1. This synergy results from both factors' presence in Sufficient Cause II, although "B" exerts a separate effect in Sufficient Cause I, and "F" in Sufficient Cause III. In a population where factor "H" were absent, "B" and "F" would exert their effects only through Sufficient Causes I and III, because the absence of "H" would eliminate any disease from Sufficient Cause II. Thus, "B" and "F" would be completely independent. In another population in which factor "C" were absent, all disease would result from Sufficient Cause II, and "B" and "F" would be fully synergistic.

An example of incomplete synergy would be the mixture of independent and interactive effects that alcohol consumption and smoking have on risk of mouth and pharynx cancer (7). Evidence suggests that either of these factors will increase cancer risk in the absence of the other, but the combined effect of both exceeds the sum of the individual effects. This would suggest at least three different causal complexes, one involving alcohol but not smoking, one involving smoking but not alcohol, and one involving both. (Some scientists have argued that alcohol in the absence of smoking has no effect, which leads to a model with two sufficient causes, one with smoking but not alcohol, and another with both smoking and alcohol.)

Discussion

The conceptual scheme in figure 1 could be modified to permit greater complexity. Antagonistic causes might be included as a part of causal chains which lead to an element being included in a sufficient cause; the joint action of the components of a sufficient cause could be considered an effect with its own constellations of sufficient causes. Alternatively, the different "pies" in figure 1 might be constructed as a sequence of causal events, with branching pathways representing those segments of pies which differ, necessary causes being common to all pathways, etc. The possibilities for adapting this framework to virtually any complicated causal relationship reinforce its utility as an intuitive base for causal thinking.

The model also provides a conceptual framework for the understanding of *latent period*, the time interval between the action of a (component) cause and the manifestation of disease. When a component cause "acts" or occurs, the other components needed to complete a sufficient cause may not be on hand. If the component cause of interest has a long-lasting "effect", as time passes the other complementary components may add their "effects" and gradually complete the sufficient cause. At the point in time at which the sufficient cause is completed, the disease process is set in motion, though usually not yet manifest. The latent period is the interval during which a sufficient cause accumulates plus the time it takes for the disease to become manifest. Early recognition could theoretically reduce the latent period to coincide with the interval during which the sufficient cause accumulates.

For some diseases, such as rabies, virtually the entire period from internalizing the viral agent to the occurrence of symptoms represents time during which the disease becomes manifest, because "internalizing" the viral agent is effectively a sufficient cause (a qualification would be that interventive treatment during the "incubation" period might prevent clinical manifestation). On the other hand, the larger part of the 10–20-year latent period for the development of vaginal cancer after *in utero*

PM3003572249

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001

exposure to diethylstilbesterol probably represents accumulation of a sufficient cause (which might ordinarily be completed at menarche or during adolescence) with only a few years for the disease to become manifest after the sufficient cause is complete. The term *incubation period* has often been applied to the period between the accumulation of a sufficient cause and the time at which disease becomes manifest. Because reversal of disease may be possible during the preclinical development, as in the example of rabies treatment, or catalysts may act to shorten the incubation period of diseases which might otherwise fester in a preclinical state interminably (growth enhancers for neoplasms are an example), it might be better to view the latent period as solely the accumulation of a sufficient cause, components of which might include developmental catalysts and/or the lack of interventive treatment.

Chronic exposures make it more difficult to quantify, and, indeed, to conceptualize latent period; different doses of an exposure accumulated over time may give rise to different risks. Such situations may be viewed as reflecting a set of different sufficient causes, each with a different dose of the exposure as a component cause. Small doses would presumably require a more complex set of complementary component causes to complete the sufficient cause than large doses. This extension of the causal model accommodates the description of dose-response relationships, and provides a basis for the common finding that for many carcinogens the dose is related directly to risk and inversely to latent period.

This presentation does not treat the subtle issues which arise in arriving at a workable definition of disease. Such issues obviously pertain to a full consideration of causes, inasmuch as disease definition may be based on experiential criteria (8), but these considerations go beyond the scope of this paper.

It might be argued that the scheme presented here is superficial because the occurrence of disease in any individual involves a collection of component causes which constitute a sufficient cause that is unique, by its complexity. Individually unique sufficient causes, however, would detract equally from all generalized causal models. Furthermore, despite individual distinctions it seems likely that there would be broad similarities in the components of sufficient causes for different individuals.

References

1. Elandt-Johnson RC: Definition of rates: Some remarks on their use and misuse. Am J Epidemiol 102:267–271, 1975
2. Miettinen OS: Confounding and effect modification. Am J Epidemiol 100:350–353, 1974
3. Rothman KJ: Synergy and antagonism in cause-effect relationships. Am J Epidemiol 99:385–388, 1974
4. Miettinen OS: Proportion of disease caused or prevented by a given exposure, trait, or intervention. Am J Epidemiol 99:325–332, 1974
5. MacMahon B: Gene-environment interaction in human disease. J Psychiat Res 6 (supp 1):393–402, 1968
6. Rothman KJ: Estimation of synergy and antagonism. Am J Epidemiol 103:506–511, 1976
7. Rothman KJ, Keller AZ: The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. J. Chronic Dis 25:711–716, 1972
8. MacMahon B, Pugh TF: Epidemiology: Principles and Methods. Boston, Little, Brown and Company, 1970

PM3003572250

Source: https://www.industrydocuments.ucsf.edu/docs/lsmj0001